A Robust Mel-Scale Subband Voice Activity Detector for a Car Platform

A. Álvarez, R. Martínez, P. Gómez, V. Nieto, and V. Rodellar

Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Facultad de Informática, Universidad Politécnica de Madrid, Madrid, SPAIN
pedro@pino.datsi.fi.upm.es

Abstract

Voice-controlled devices provide a smart solution to operate add-on appliances in a car. Although, speech recognition appears as a key technology to produce useful end-user interfaces, the amount of acoustic disturbances existing in automotive platforms usually prevents satisfactory results. In most of the cases, noise reduction techniques involving a Voice Activity Detector (VAD) are required. Through this paper, a robust method for speech detection under the influence of noise and reverberation in an automobile environment is proposed. This method determines a consistent speech/non-speech discrimination by means of a set of Order-Statistics Filters (OSFs) applied to the log-energies associated to a mel-scale based subband division. The paper also includes an extensive performance evaluation of the algorithm using AURORA3 database recordings. According to our simulation results, the proposed algorithm shows on average a significantly better performance than standard VADs such as ITU-G.729B, GSM-AMR or ETSI-AFE, and other recently reported methods.

Index Terms: voice activity detection, robust speech recognition

10.21437/Interspeech.2007-92

1. Introduction

Speech uttered inside a car is perturbed by an appreciable amount of noise and reverberation. Environmental conditions produce a deep impact in speech processing applications affecting the overall system performance. Numerous techniques have been derived to palliate their harmful effects on performance. Traditionally, noise reduction techniques have mainly utilized a one-microphone structure, with or without any hypothesis on the noise/speech distribution [1], [2], [3]. These methods are based on the signal-to-noise ratio (SNR) estimation and make use of the speech intermittence and noise stationarity hypothesis.

Precise noise statistics required by cancellation algorithms are usually supported by a speech/non-speech decision. Therefore, the contribution of a reliable Voice Activity Detector (VAD) is involved. A VAD, when applied, plays a central role in the speech enhancement task as its accuracy dramatically affects the noise suppression level and amount of speech distortion that occurs. Besides, the related speech/non-speech classification procedure is not trivial as most part of algorithms fail when the level of background interference increases.

A representative set of the reported VADs formulate the decision rule on a frame by frame basis using smoothed estimations of the speech and noise present in the signal [4], [5]. Speech and noise measures are usually modeled as a Gaussian distribution (eg. [4], [6]), Laplacian [7] or even both [5]. Besides, VAD robustness may be improved by using speech signal long-term information [8] and nonlinear filtering as Order Statistic Filters (OSFs) [6].

Through this paper, a VAD based on the use of a set of Order Statistic Filters applied to the log-energies calculated for a subband division on the mel scale, is described.

2. OSF based Voice Activity Detector

This section describes the MSSQ (Mel-Scale Subband Quantile) VAD. This method addresses the use of OSFs [9] for robust speech and noise estimation. Calculations are done concurrently over a spectrum partition into mel-scale subbands. Subband SNRs are tested individually on a noise-dependent threshold function. Final speech/non-speech assessment is adopted by the aggregation of individual decisions achieved on every subband.

2.1. Subband log-energy calculation

The noisy speech signal *x(n)* is segmented in 64 ms frames with a 16 ms window shift and transformed into the frequency domain using the short-time Discrete Fourier Transform.

Let *X(m,k)* be the spectrum magnitude for the k^{th} FFT bin *(k=0, 1, ..., L-1)* at frame *m*. The log-energies *E(m,b)* for the m^{th} frame and the b^{th} band *(b= 0,1, ..., B-1)* are calculated as follows

$$E(m,b) = 10.0 \cdot \log_{10}\left(\frac{B}{L/2} \sum_{k=k_b}^{k_{b+1}-1} |X(m,k)|^2 \right) \quad (1)$$

where k_b is the first FFT bin associated to the b^{th} subband. Subbands are established by means of a uniform division on the mel scale. The relation between both scales is defined by

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \quad (2)$$

2.2. Speech and noise estimations

The subband log-energies are used in conjunction with two OSFs to calculate smoothed out estimations of the speech *S(m,b)* and noise *N(m,b)*, respectively. These estimations exploit long-term spectral information over a *N*-frame neighborhood in order to enhance their robustness and provide a simple hangover mechanism.

$$S(m,b) = Q_\alpha[E(m-N,b),...,E(m,b),...,E(m+N,b)] \quad (3)$$

$$N(m,b) = Q_\beta[E(m-N,b),...,E(m,b),...,E(m+N,b)] \quad (4)$$

where $Q_x[Y]$ represents the *x-quantile* of a serie *Y*.

Quantile values linked to speech and non-speech estimations are $\alpha = 0.9$ and $\beta = 0.3$, respectively. On the other hand, practical values of *N* are in the range *[2, 6]*.

A SNR function is computed as follows

$$SNR(m,b) = S(m,b) - N'(m,b) \quad (5)$$

where *N'(m,b)* is derived from *N(m,b)* by means of a 1st order IIR filter, that is

$$N'(m+1,b) = \lambda N'(m,b) + (1-\lambda)N(m,b) \quad (6)$$

where $\lambda=0.95$.

It is important to notice that noise estimations are only updated during non-speech periods (depending on the previous VAD decision). Finally, for algorithm initialization purposes, the first frame is assumed to be non-speech

$$N(0,b) = E(0,b) \quad b = 0,1,...,B-1 \quad (7)$$

2.3. SNR threshold function definition

A set of threshold functions $T_H(m,b)$ are defined dependent on the estimated noise level *N'(m,b)*. As it may be seen, $T_H(m,b)$ decreases linearly with increasing noise levels

$$T_H(m,b) = \eta_{30} - \left(\frac{\eta_{30}-\eta_{120}}{120-30}\right)N''(m,b) \quad (8)$$

where η_{30} and η_{120} are the threshold values associated to noise levels of 30dB and 120dB, respectively. Besides, *N''(m,b)* is simply *N'(m,b)* but limited to the range *[30, 120]*

$$N''(m,b) = min\{max[N''(m,b),30],120\} \quad (9)$$

Speech detection is carried out for every subband separately

$$\begin{aligned} V(m,b) &= 1 \quad if\ SNR(m,b) > T_H(m,b) \\ V(m,b) &= 0 \quad else \end{aligned} \quad (10)$$

The overall VAD decision is then elaborated from individual subband decisions *V(m,b)*

$$VAD(m) = \sum_{b=b_{start}}^{B-1} V(m,b) \quad (11)$$

where $0 \le b_{start} \le B-1$ is the first subband being considered in order to determine the presence of speech at frame *m*.

The use of b_{start} higher than zero results in an increased cut-off frequency for the speech analysis. The main goal is to eliminate subbands in which SNR testing is less reliable as car noise in not white and it is mainly concentrated in low frequencies [10].

As a final step, if *VAD(m)* is greater than zero, the actual frame is classified as speech (H_1), otherwise it is classified as non-speech (H_0)

$$VAD(m) \underset{H_0}{\overset{H_1}{\gtrless}} 0 \quad (12)$$

2.4. Hangover scheme

The system hangover scheme is based on a simple two state automaton (S_0 and S_1). State S_0 implies the previous frame was considered as non-speech and S_1 just the opposite. As mentioned before, noise adaptation in (6) is only carried out whether next state is going to be S_0. Apart from *N'(m,b)* calculation, the current state also affects the threshold function. If current state is S_1 (previous frame was labelled as speech) η_{30} and η_{120} are substituted by η'_{30} and η'_{120} in (8). The four thresholds are shown in Figure 1. As it may be seen, speech periods are associated a lower threshold value, also linearly related to actual noise level. The idea behind is to prevent clipping of weak speech tails.

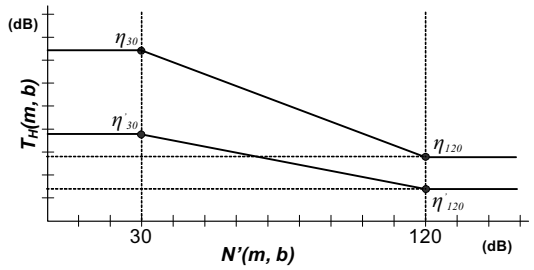

Figure 1. *Threshold functions established for non-speech periods (upper line) and speech periods (lower line), respectively.*

3. Results and discussion

Several experiments were conducted to evaluate the performance of the proposed method (WSSQ) against other VADs. The analysis is focused on the classification capabilities at different SNR levels. Misclassification errors were studied at different SNR levels by means of the receiver operating characteristics (ROC) curves. Also, the influence of the proposed VAD on speech recognition tasks was assessed. Standard VADs as G.729-annex B [11], AMR [12] and AFE [13], as well as methods reported by Sohn [4] and Ramírez [6] were used for reference.

3.1. Analysis of the ROC curves

The speech material used for the analysis is the AURORA3 subset of SpeechDat-Car Danish, Finnish, German and Spanish databases [14]. This corpus contains realizations of connected digits uttered in a realistic automobile environment by more than several hundred speakers. AURORA3 includes a close-talking channel (*ch0*) and far distant microphone recordings (*ch1*). Train and test sets are subdivided into three categories regarding the noise level. Finally, average SNR values are between 25 dB and 5 dB.

The whole corpus was partially hand-labeled on the close talking microphone to obtain the speech/non-speech hit rates, *HR1* and *HR0* respectively.

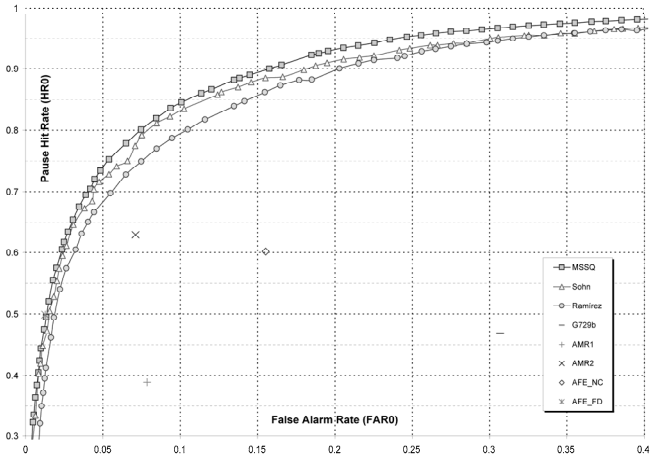

Figure 2. *ROC curves for the distant microphone (ch1) and the three driving conditions.*

Database	Train/Test	Word Accuracy							
		No VAD	*Standard VADs*				*Other reported VADs*		MSSQ
			G.729B	AMR1	AMR2	AFE-FD	Ramírez	Sohn	
Danish	HM	26.18 %	28.66 %	43.53 %	52.75 %	39.51 %	46.65 %	45.85 %	**51.88 %**
	MM	34.34 %	28.43 %	54.12 %	54.12 %	54.12 %	58.52 %	67.87 %	**61.21 %**
	WM	76.14 %	74.16 %	85.41 %	85.72 %	85.32 %	86.13 %	87.79 %	**87.30 %**
	Average	45.55 %	43.75 %	61.02 %	64.20 %	59.65 %	63.77 %	67.17 %	**66.80 %**
Finnish	HM	34.06 %	43.53 %	54.38 %	51.73 %	34.13 %	51.86 %	48.66 %	**54.95 %**
	MM	68.13 %	61.08 %	75.51 %	75.85 %	73.12 %	77.14 %	80.78 %	**79.69 %**
	WM	88.09 %	88.72 %	94.38 %	94.62 %	92.61 %	93.21 %	92.19 %	**93.79 %**
	Average	63.43 %	64.44 %	74.76 %	74.07 %	66.62 %	74.07 %	73.88 %	**76.14 %**
German	HM	69.98 %	69.66 %	73.22 %	73.22 %	74.19 %	73.08 %	73.15 %	**76.41 %**
	MM	77.67 %	66.84 %	78.70 %	77.96 %	78.26 %	81.98 %	75.56 %	**77.90 %**
	WM	90.14 %	86.36 %	90.86 %	90.84 %	91.67 %	91.91 %	90.21 %	**91.46 %**
	Average	79.26 %	74.29 %	80.93 %	80.67 %	81.37 %	82.32 %	79.64 %	**81.92 %**
Spanish	HM	53.83 %	53.71 %	60.81 %	70.26 %	64.54 %	67.22 %	67.73 %	**70.02 %**
	MM	68.59 %	71.34 %	84.64 %	85.00 %	85.82 %	86.37 %	84.58 %	**87.65 %**
	WM	86.17 %	88.84 %	93.57 %	94.22 %	94.25 %	94.84 %	93.96 %	**95.20 %**
	Average	69.53 %	71.30 %	79.67 %	83.16 %	81.54 %	82.81 %	82.09 %	**84.29 %**
Average		**64.44 %**	**63.45 %**	**74.10 %**	**75.53 %**	**72.30 %**	**75.74 %**	**75.70 %**	**77.29 %**

Table 1. *Recognition results for the AURORA3 SpeechDat Car databases (Framework in Figure 3.a).*

Database	Train/Test	Word Accuracy							
		No VAD	*Standard VADs*				*Other reported VADs*		MSSQ
			G.729B	AMR1	AMR2	AFE-FD	Ramírez	Sohn	
Danish	HM	57.28 %	40.86 %	62.85 %	78.55 %	76.54 %	79.16 %	78.56 %	**79.78 %**
	MM	51.79 %	44.78 %	57.63 %	65.24 %	63.05 %	70.47 %	74.14 %	**67.77 %**
	WM	81.47 %	79.60 %	89.34 %	90.76 %	90.57 %	91.28 %	91.21 %	**91.57 %**
	Average	63.51 %	55.08 %	69.94 %	78.18 %	76.72 %	80.30 %	81.30 %	**79.71 %**
Finnish	HM	57.95 %	67.63 %	85.41 %	91.59 %	84.42 %	86.03 %	89.40 %	**90.60 %**
	MM	75.03 %	68.67 %	84.88 %	81.87 %	80.23 %	79.86 %	85.63 %	**85.43 %**
	WM	96.15 %	90.39 %	96.00 %	96.83 %	95.95 %	95.32 %	95.52 %	**96.31 %**
	Average	76.38 %	75.56 %	88.76 %	90.10 %	86.87 %	87.07 %	90.18 %	**90.78 %**
German	HM	88.07 %	78.77 %	87.65 %	89.87 %	91.30 %	87.50 %	86.60 %	**88.11 %**
	MM	85.43 %	71.52 %	86.68 %	86.29 %	89.75 %	82.49 %	81.31 %	**85.89 %**
	WM	93.45 %	88.38 %	94.23 %	94.15 %	95.15 %	94.01 %	93.30 %	**94.47 %**
	Average	88.98 %	79.56 %	89.52 %	90.10 %	92.07 %	88.00 %	87.07 %	**89.49 %**
Spanish	HM	63.85 %	70.95 %	79.79 %	92.06 %	91.76 %	91.19 %	90.09 %	**92.33 %**
	MM	82.94 %	79.24 %	87.89 %	93.13 %	92.58 %	92.12 %	91.34 %	**93.62 %**
	WM	90.55 %	91.76 %	95.64 %	96.66 %	96.40 %	96.57 %	96.22 %	**97.16 %**
	Average	79.11 %	80.65 %	87.77 %	93.95 %	93.58 %	93.29 %	92.55 %	**94.37 %**
Average		**77.00 %**	**72.71 %**	**84.00 %**	**88.08 %**	**87.31 %**	**87.17 %**	**87.78 %**	**88.59 %**

Table 2. *Recognition results for the AURORA3 SpeechDat Car databases (Framework in Figure 3.b).*

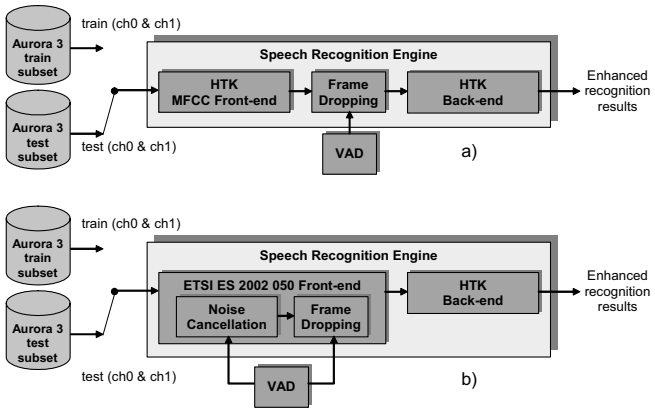

Figure 3. *a) Baseline HTK-based non-robust speech recognizer incorporating a VAD. b) ETSI robust front-end controlled by a single VAD.*

Figure 2 shows the trade-off between speech pause hit rate (*HR0*) and false alarm rate *FAR0* (*FAR0= 1.0-HR1*). The proposed VAD works with lower alarm rate and higher speech-pause hit rate when compared to standard methods, especially over G.729. The only algorithm presenting a similar working point is the AFE-NC (AFE VAD devoted to the Wiener filtering noise cancellation task). However, that situation is far from optimal, since *HR0* is around 50%, thus leading to an extremely conservative behavior in detecting speech pauses. The MSSQ VAD also outperforms Sohn and Ramírez VADs although the improvement is reduced in the area corresponding to high *HR1* values.

3.2. VAD influence on an ASR system

Several speech recognition systems were tested using two different frameworks (see Figure 3), both based on HTK (Hidden Markov Toolkit) package [15]. The first system uses a baseline MFCC front-end (12 cepstral coefficients, logarithmic frame energy plus delta and delta-delta coefficients) with the inclusion of a frame dropping mechanism in order to increase recognition rates.

The second framework is built with ETSI AURORA robust front-end for DSR [13]. In this case, a single VAD controls the Wiener filtering and frame dropping stages whereas the standard system has two different endpoint detectors (AFE-NC and AFE-FD). This front-end extracts the same 39 parameter set already mentioned. For the AURORA3 databases, the so called high-mismatch (HM), medium-mismatch (MM), and well-matched (WM) training/testing conditions are used. The task consists of recognizing connected digits which are modeled as a whole word Hidden Markov Model with the following configuration.

Working point for Sohn, Ramíred and the proposed VAD (MSSQ) is chosen so that it improves the overall recognition accuracy. MSSQ method is configured with the following parameters: B=15, b_{start}=3 (subbands 0, 1 and 2 are removed), N=4, η_{30}=15dB, η_{120}=3.5dB, η'_{30} =9dB and, η'_{120} =2.5dB.

Table 1 shows the recognition performance achieved by the different algorithms for the ASR architecture shown in Figure 3.a. As it may be seen, the proposed MSSQ algorithm outperforms all the VADs used for reference, being the improvements more relevant when compared with the majority of the standard methods. Moreover, G.729 produces worse results than baseline system (No VAD) in which all frames are passed to the recognizer back-end. AFE-FD exhibits a reduced accuracy produced by the fact that, there is not a prior noise cancellation module available in this case. The only exception to the rule appears in the German database where the working-point, producing a high *HR1* but at a cost of a poor detection of non-speech periods is clearly honored, as the high scores achieved by the baseline VAD indicate. Finally, GSM-AMR2, Sohn and Ramírez methods provide similar recognition rates but behind MSSQ.

Table 2 presents the experimental results for the framework in Figure 3.b. An improved accuracy as a result of the application of a robust front-end is clearly reported for all the algorithms using the same parameter configuration. In general VADs share a similar trend in both frameworks. The method proposed in this paper (MSSQ) attains, as in the previous test the best overall figures. However, the word error rate (WER) reduction related to the original ETSI front-end (with the AFE-VAD) is reduced, mainly because the Wiener filtering and frame dropping stages (AFE-NC and AFE-FD, respectively) are tuned to VADs with a conservative working point. Besides, AFE-VAD also benefits from the use of a double VAD scheme in the second framework. However, the word accuracy of the ETSI-AFE method is still behind AMR2, Sohn and, MSSQ.

4. Conclusions

This paper presents a robust Voice Activity Detector (VAD) devoted to work in voice-controlled devices on cars. The current approach is based on a effective endpoint detection algorithm employing order statistic filters for the estimation of noise and speech. The formulation of the decision rule is built upon a spectral subband analysis. A uniform division on the mel-scale contributes to clear improvements especially for low SNR scenarios. This work includes several ASR evaluations carried out with real data taken form the AURORA3 database. The proposed algorithm outperforms G.729, AMR1, AMR2 and AFE standards and, other VADs among the best reported, in speech/non speech detection capabilities.

5. Acknowledgements

This work is being funded by grants TIC2003-08756 and TEC2006-12887-C02-01/02 from Plan Nacional de I+D+i, Ministry of Education and Science, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

6. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, 1979.

[2] R. J. McAulay, M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 2, pp. 137- 145, 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109- 1121, 1984.

[4] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection", IEEE Signal Processing Letters, vol. 16, no. 1, pp. 1-3, 1999.

[5] S. Gazor, W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model", IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 498- 505, 2003.

[6] J. Ramirez, J.C. Segura, C. Benitez, A. de la Torre, A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 13, no. 6, pp. 1119-1129, 2005.

[7] J.H. Chang, N.S. Kim, "Voice activity detection based on complex Laplacian model", Electronics Letters, vol. 39, no. 7, pp. 632- 634, 2003.

[8] J. Ramírez, J.C. Segura, C. Benítez, A. de la Torre, A. Rubio, "Efficient voice activity detection algorithms using long-term speech information", Speech Communication, no. 42, pp. 271-287, 2004.

[9] H.D. Tagare, de Figueiredo, R.J.P., "Order Filters", Proceedings of the IEEE, vol. 73, no. 1, pp.163-165, 1985.

[10] C.E. Mokbel, F.A. Chollet, "Automatic Word Recognition in Cars", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 5, September 1995, pp. 346-356.

[11] ITU, "A Silence Compression Scheme for G.729 optimized for terminals conforming to recommendation V.70", ITU-T Recommendation G.729-Annex B, 1996.

[12] ETSI, "Voice Activity Detector (VAD) for Adaptive MultiRate (AMR) speech traffic channels", ETSI EN 301 708, 1999.

[13] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Advanced front-end feature extraction algorithm", ETSI ES 2002 050, v1.1.5, January 2007.

[14] A. Moreno, et al., "SPEECHDAT-CAR: A Large Speech Database for Automotive Environments", Proc. of 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 2000, paper 373.

[15] S. Young, et al., The HTK Book (for HTK Version 3.4), Cambridge University, 2005.